AD694745

ON LIMITS IN COMPUTING POWER

Willis H. Ware

October 1969

OCT 20 1969

...has been approved
...sale; its
distribution is unlimited.

P-4208

Reproduced by the
CLEARINGHOUSE
for Federal Scientific & Technical
Information Springfield Va. 22151

9

ON LIMITS IN COMPUTING POWER

Willis H. Ware*

The Rand Corporation, Santa Monica, California

At one time or another you have probably all heard of the growth figures quoted for the computing industry in the double decade of 1955-1975; these figures are part history and part extrapolation, but to the extent that history has progressed since the estimates were made, the extrapolations are valid. In these twenty years the size of the computer has decreased 10,000-fold for equal computational capability. The unit cost of calculation is down by the startling figure of 200,000-fold, while speed has increased 40,000-fold. Also, there has been an explosive growth of installed capacity, which over the double decade of 1955-1975 has increased 160,000-fold. The '70s have been extensively analyzed and projected, and by 1975 or so machines ought to operate close to 10^9 operations per second. This morning I thought it would be more exciting to move on into the '80s to see what limits might set a ceiling on computational capability. These thoughts do not reflect original research on my part; rather I have tried to extrapolate from the work of others. There is not universal agreement about the arguments on which I draw, so my conclusions must be considered as "ballpark" guidelines.

Although I recognize that we can conceivably get increasing capability from software improvements, or from better numerical analytic techniques and better mathematics, I want to avoid these issues today and talk about (1) the hardware, particularly with respect to component

---

*This paper is not a formal research product produced for a client of The Rand Corporation. Thus, it is not the view, official opinion, or policy of Rand nor of any of its governmental or private research sponsors. However, it does reflect the general participation of the author in the Rand research program and his immersion in the Rand research environment. Papers are published by The Rand Corporation to support the professional activities of its staff members.

This paper was presented at the 1969 Meteorological Technical Exchange Conference, U.S. Air Force Academy, Colorado, 14-17 July 1969.

speeds and the limitations imposed by the laws of physics; (2) the logical arrangements used to implement arithmetic; and (3) the overall machine architecture.

First, let me address the question of arithmetic logic. Several years ago Winograd* of the IBM Research Laboratory undertook to investigate the maximum speed with which arithmetic can be done. The parameters of the problem are obvious: the length of the numbers to be handled, the so-called fan-in of the logic element (the number of signals one logical element can accept), the base of the number system (binary or decimal), and the delay time for the logic element. Winograd was able to establish a formula that predicts the absolute minimum time in which addition can be done. In order to achieve such a minimum, numbers will have to be represented in remainder, or residue form, rather than in conventional positional notation. Winograd also addressed the question of multiplication, and he found that in some cases multiplication can actually be done more rapidly than addition. Again, the numbers have to be expressed in a special way. The stickler is that addition and multiplication require different special representations.

Therefore, it would appear that an inevitable compromise between addition speed and multiplication speed will have to stand. At this time, it does not seem worthwhile to design a machine in which numbers have two special representations for the sole purpose of speeding up arithmetic.

Winograd's analysis brought an additional point to light. Such operations as overflow determination, as well as any operation that depends on it such as COMPARE, cannot be speeded up. As we know computation today, overflow indication is essential; and, therefore, representation of numbers in such special ways as residue form is not an attractive option.

---

*S. Winograd, "On the Time Required To Perform Addition," *J. Assoc. Comput. Mach.*, Vol. 12, No. 2, April 1965, pp. 277-285; idem, "On the Time Required To Perform Multiplication," *J. Assoc. Comput. Mach.*, Vol. 14, No. 4, October 1967, pp. 793-802; J. F. Brennan, "The Fastest Time of Addition and Multiplication," *IBM Research Reports*, Vol. 4, No. 1, 1968 (a digest of the two Winograd papers).

In fact, well-designed contemporary machines do addition at roughly 60 to 80 percent of the Winograd limit. When one considers that machine designers have had no metrics to guide them, this is a remarkable achievement. On the other hand, multiplication is running a quarter to a third of its maximum rate; but even so, it does not yet look worthwhile to represent factors in special form to enhance multiplication speed. The significant conclusion is that any big gains one can anticipate in computers will not come from the logical arrangements to implement arithmetic. We might squeeze 10 to 20 percent in addition and/or 50 percent in multiplication, but there will not be orders-of-magnitude improvement from the logical implementation of the arithmetic processes.

If we examine the arithmetic unit (CPU) of contemporary machines and ask how efficiently it is used, we discover that it is idle a substantial amount of time; typically CPU utilization is about 50 percent, and it can be lower. There are bottlenecks in the internal information flow; e.g., the arithmetic unit is frequently waiting for the memory or for the magnetic tapes. Thus, the efficiency of utilization is not as high as desired. More sophisticated designs to appear in the early '70s will provide a steadier flow of information to the arithmetic unit but there is only a factor of roughly 2 or 3 to be gained. Thus, it follows that any machine that performs only a single arithmetic operation at a time is within a factor of 3 to 5* of the end of the line. If such a machine is to improve any more, it must utilize faster components. It also follows that in the early or mid-'70s we will have to turn increasingly for super machines to the multistream concept such as represented by Illiac IV or some of the pipeline-machine processors now in design. In such machines, a large number of arithmetic operations are in process concurrently.

Let me next turn to component technology. Consider the conclusion obtained by Bremermann, of the University of California at Berkeley, and also formulated by Marko.** Bremermann has published twice the

---

*This is a composite figure including a factor of 2 to 3 for CPU efficiency and one of 1½ for pushing arithmetic to the Winograd limits.

**H. Marko, "Physikalische und biologische Grenzen der Informationsübermittlung," *Kybernetik*, Vol. 2, No. 6, October 1965, pp. 274-284.

conclusion that no data processing system, be it artificial or living, can process more than c^2/h bits per second per gram, c being the velocity of light and h being the Planck constant. In one publication* he bases his argument on quantum-mechanical principles, and in the other** he bases it on thermodynamical principles. This result can be challenged but if his limit stands, it has interesting consequences. If we insert values, c^2/h becomes $2 \cdot 10^{47}$ bits per second per gram. Hypothesizing a computer of the mass of the earth and of the age of the earth, we find that such a machine could have processed only 10^{93} bits in its lifetime. This appears to be an enormous number, but in reality it is small compared with some of the problems people are discussing. For example, the number of move sequences in a chess game is approximately 10^{120}; a straightforward attack on the problem would require a capability far beyond that of our earth-size, earth-age computer. Similarly, a picture of 100 by 100 cells, each of which can be black or white, contains 10^{3000} different patterns. No doubt, some of the meteorological problems that are under discussion cannot be dealt with by a routine, brute-force, head-on collision with a super computer. The mathematical analysis will have to be very ingenious to bring such problems within range of attack.

Of course, we are presently nowhere near this limit, so let's discuss individual components. If we wish to store information, we need a device that has two potential wells separated by a barrier;*** one potential well corresponds to binary zero, the other to binary one. To change the state of the device, energy must be inserted to move over the barrier, and if the device is expected to stay in the new

---

*H. J. Bremermann, "Optimization through Evolution and Recombination," in *Self-Organizing Systems 1962*, Spartan Books, Washington, D.C., 1962, pp. 93-106.

**Idem, "Quantum Noise and Information," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. IV, University of California Press, Berkeley and Los Angeles, California, 1967, pp. 15-20.

***R. Landauer, "Irreversibility and Heat Generation in the Computing Process," *IBM J. Res. and Develop.*, Vol. 5, No. 3, July 1961, pp. 183-191.

state, the energy must be removed when the new state is reached. The random energy of motion is of the order of kT, so that if the device is to behave reliably, the barrier must be a few kT high.* Thus, the minimum energy that needs to be expended per information event is of the order of a few kT, but the significant conclusion is that energy must be dissipated as heat. Unavoidably, computing involves dissipation of heat; there is no way to circumvent the problem if we are to build a computing device that is to be reliable.

The next consideration is that computation destroys information. Thus, it is a nonlinear process and depends for implementation on logical functions that are also nonlinear and that depend in turn on nonlinear phenomena for practical realization.** Nonlinearity of electronic components also contributes to the practical problems of the computer; e.g., restandardization of signals, fabrication tolerances, and noise rejection. In the present solid-state technology, signals of a quarter volt or so are necessary to maintain the nonlinearity of the *p-n* junction in semiconductors and to absorb fabrication tolerances. This is not a limitation of a theoretical law of physics but rather a state-of-the-art limitation. It is anticipated that improved devices--not of the semiconductor type--will be found that maintain nonlinear behavior with signals 10 to 20 times smaller. Because a signal of finite voltage amplitude is inevitably necessary, capacitance-charging problems set a final limit to the speed at which a component can operate.

We consider now the velocity of light that is an absolute limit on the speed with which information can move. If we want to build fast computers, we must build small ones; and we will have to package them densely. However, small size and dense packaging are inconsistent with heat dissipation. The heat dissipation problem appears to be a more serious constraint than any others now visible.

Let me suggest the scale of the problem. In modern-day transistors,

---

*k is Boltzmann's constant, and T is the Kelvin temperature.

**M. J. Freiser and P. M. Marcus, "A Survey of Some Physical Limitations on Computer Elements," *IBM Research Note*, RC 2283, November 14, 1968, pp. 7-8. This article is to be published in *IEEE Transactions on Magnetics*, June 1969.

the power density inside the transistor at the *p-n* junction is of the order of thousands of watts per square centimeter.* In contrast, the maximum heat transfer to fluids at approximately room temperature is about 100 watts per square centimeter. There is a factor of 10 or so that somehow has to be accommodated. Obviously, we need to spread the heat over a large enough surface so that it can be transferred to a fluid. Thus, the mismatch between internal working power densities and external fluid absorption power densities is a major constraint on the minimum size of components, and hence on the speed with which they can operate.

Where are we today? Thin superconducting films can be switched in research environments at about 10^{-10} seconds.** The capacitance-charging problem in semiconductor junctions sets limits at about 10^{-11} seconds. The time for carriers to drift across the base of a transistor, given the technology that we can project for making very small base widths, is of the order of 10^{-10}. We can switch magnetic films in about 10^{-9} seconds and magnetic cores in about 10^{-7} seconds. These are all state-of-the-art limitations. Interestingly enough, except for the core, they are all in the general neighborhood of 10^{-10}.

Where do the laws of physics impose limitations? The cooling problem sets a practical limit on switching time at about 10^{-11} seconds.*** If clever engineering can solve this problem, we look forward to speeds of 10^{-12} to 10^{-13} seconds. Of course, there is a fundamental limit at 10^{-15} seconds due to indeterminacy. Present research results are not very far from what appear to be absolute limitations, and thus, we should anticipate computing elements that will switch information states in about 10^{-11} to 10^{-12} seconds.

Where is the state of the art today? Present production devices switch in about 10^{-8} seconds, and present research items switch in about 10^{-9} to 10^{-10} seconds. Depending on what one wishes to use as

---

*Ibid., pp. 9-10.

**Ibid., pp. 12-14.

***Eugene G. Fubini, private communication to the author. See also R. W. Keyes, "Physical Problems and Limits in Computer Logic," *IEEE Spectrum*, Vol. 6, No. 5, May 1969, pp. 36-45.

a "practical" upper limit for component speed and what one chooses as the present-day state of the art in research, there is a factor of at least 100 (from 10^{-9} to 10^{-11} seconds) yet to be realized from component technology speed. If we can push beyond 10^{-11} to 10^{-12} seconds, we will have a 1000-fold improvement, but with present understanding of the theoretical and practical limits, it does not appear that factors beyond a few thousand will ever be achieved. Even the minimum improvement of 100-fold is an impressive future to contemplate.

Machine architecture is a difficult subject to treat. Most of the experience in the computing field has been with machines executing a single instruction stream, doing one arithmetic operation at a time, and organized internally so that the arithmetic unit is maximally utilized. Experience with multistream machines is limited. Taking into account the estimates of the Illiac IV machine, which is about as multistream as now envisioned, and discounting somewhat the hopes of its builders, we may be able to achieve an increase of 100-fold (as opposed to projected factors of many hundreds). This factor depends strongly on how much of the problem is inherently serial; 100-fold implies that 1 percent of the problem is serial. Combining this with the smallest factor of 10^2 that component technology has yet to go, we may eventually get as much as 10^4, or a 10,000-fold increase in raw computing power. If problems prove to be "more parallel" than we think, or if we do push technology even further, the overall improvement could move toward 100,000-fold. This is an even more impressive future to anticipate for the environmental problems with which you are concerned.

Any such mammoth machine would be very special and probably warranted only for the problems that could exploit it. The commercial industry is not likely to build such a machine unless a large market appears. Construction, at least for the first one, will no doubt have to be funded separately, and if your problems need such computer power, be prepared to finance the development of such machines and to dig deeply into your budgets. I won't project the development cost, but it will be substantial although probably less than a large particle accelerator.

Acknowledgments: I appreciate the critical review and very helpful comments from M. Warshaw and J. C. Shaw of The Rand Corporation, Dr. Rolf Landauer of IBM, and Dr. Eugene Fubini, consultant.